USE OF CLASSICAL STATISTICS in ...

DERIVING and EVALUATING CER's

A PRESENTATION TO
OPERATIONS RESEARCH SOCIETY OF AMERICA

Durham, North Carolina

by

1st Lt. Charles Graver

Office of Assistant Secretary of Defense (Systems Analysis)

THE USE OF CLASSICAL STATISTICS IN DERIVING AND EVALUATING CERs

Operations Research Society of America
Durham, North Carolina

17 October 1966

Presentation Given By:

1st Lt. Charles Graver, USAR
OASD (Systems Analysis)
Resource Analysis

Acknowledgements

The author wishes to thank Mr. Paul Dienemann, Mr. Saul Hoch, Mr. David Lovejoy and Mr. Harry Piccariello for their valuable comments during the preparation of the written form of this presentation.

Table of Contents

INTRODUCTION

Regression theory is frequently used in the development of cost estimating relationships (CERs). Unfortunately there is a tendency to use this tool and the statistics that are associated with it without fully understanding either. A good understanding, however, is necessary for both the user and the reviewer.

To develop this understanding I would like to discuss two main topics. The first centers on the meaning of some of the commonly used statistics and the differences between common interval estimates. The second addresses the applicability of the usual interpretation of these statistics and interval estimates in cost analysis and the possible meanings that might be attached to them even if statistical assumptions are not fully satisfied.

This presentation will address the above through a discussion of the following topics:

1. Assumptions of the multiple linear regression model and how well they are fulfilled in the cost analysis application.
2. Least squares estimators as "best" estimators.
3. Properties of some commonly used statistics from a geometrical point of view.
4. Differences in commonly used interval estimates.

For most of the presentation a multiple linear regression model will be assumed with only passing remarks to other types of regression functions. It is assumed that the reader has some experience in using statistics. Certain terms will therefore not be defined. Any good statistics book[1] should give the definitions of terms that are unknown to the reader.

[1] Lindgren, B. W., Statistical Theory, Macmillan, New York, 1962.

THE MODEL

The usual form of the multiple linear regression model is given below.

$$Y_i = a + b_1 X_{1i} + b_2 X_{2i} + \dots + b_k X_{ki} + e_i$$

$$\text{for } i = 1, 2, \dots, n$$

For each of the n sample points, indexed by i, Y_i is the sample observation, $a + b_1 X_{1i} + b_2 X_{2i} + \dots + b_k X_{ki}$ is the value of the regression function and e_i is the error term.

Regression theory assumptions fall into two categories, those pertaining to the regression function and those pertaining to the error term. They are listed below:

Assumptions

Regression Function

1. Independent variables are non-random.
2. The regression function is a true relation.

Error Term

3. Normally distributed
4. Identically distributed
5. Mutually independent
6. Random sample.

The independent variables (not statistically independent but functionally independent) are assumed to be <u>non-random</u>, i.e., their value is not in doubt. It should be noted that they need not be just one characteristic such as thrust, but might represent a function of characteristics such as weight times speed squared. That is $X_{1i} = T_i$ and $X_{2i} = W_i(S_i)^2$ where T_i, W_i and S_i are the thrust, weight and speed of the i^{th} observation respectively.

The variables must be brought together by the regression function in such a way that a true relation is represented between the sample observations, Y_i, of cost in our case, and the characteristics that make up $X_{1i}, X_{2i}, \ldots, X_{ki}$. By true relation I mean that the sample observations and the characteristics must be related in a semi-deterministic way, that is, if the error term e_i were not present in the model, then the relationship would be deterministic. It is required in the multiple linear regression model that the regression function be linear in the parameters $a, b_1, b_2, \ldots, b_k$. Other models do exist, such as the log-linear regression model, that treat other forms of the regression function.

Turning to the error term, it is usually assumed that the e_i are normally distributed with zero mean and a common but unknown variance σ^2. Hence, they are identically distributed. It is also assumed that the e_i are mutually independent, that is, knowing the value of any e_i, say e_1, does not change the distribution on any of the other e_i.

The final assumption listed, that of a random sample, is really equivalent to the independent and identically distributed assumptions on the e_i. It is mentioned here to point out that a sample from a population is assumed.

A graphic representation of the simple linear regression model is shown in Figure 1. Note that for each value of X there is a normal density function (represented by the bell shaped curve) centered on the true regression line Y = a + b X. The densities all look the same reflecting the assumed identical distribution of the error terms, e_i.

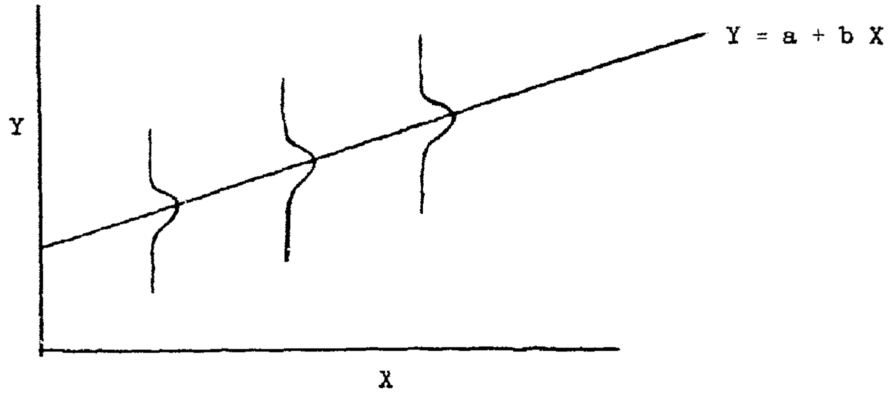


Figure 1.

How well are these assumptions met in the typical cost analysis application? In general, not very well. Let's take them one at a time and see.

Independent variables are non-random. There are two problem areas here for the cost analyst. First, unknown to the analyst, the characteristic desired may lack a common definition for all objects in the sample. For example, payload might be a desirable characteristic, but it may not be defined the same for fighters and transport aircraft.

If both are included in the sample some of the observed variation will be due to the definition of the characteristic. Secondly, for lack of anything better one might attempt to estimate some costs as functions of other costs, for example

$$C_O = a + b\ C_I$$

where C_O is the operating cost

and C_I is the investment cost.

Unfortunately the independent variable (C_I) is in general random.

The regression function is a true relation. This topic will be covered later on by another paper on this panel.[1] Let it suffice to say here that the regression function should be picked ahead of time (based perhaps on some logical or physical relationship). Allowing a machine to pick the regression function might yield a "good fitting" relationship, but one that has little to do with the real world.

Normal distribution. This assumption is the least restrictive, for if we had a large sample, it would not be necessary to worry about it. But cost analysts are not given this luxury in general. There are, however, plenty of other problem areas to concern ourselves with, so the pros and cons of the small sample normal assumption will not be discussed further in this presentation.

Identically distributed. This assumption is often violated. For example, in order to build up the sample size, bomber and fighter aircraft may be included in the same sample. If the cost that one is

[1] Yates, Edward H. and Frederic, Brad C., Cost Hypothesis: Their Development and Their Data Implications, D.R.C.

trying to predict depends on mission, then one would expect that the error term would require a different mean and/or variance assumption for the two classes of aircraft. This is because the sample should be stratified and it would be difficult to devise a regression function that could handle this stratification. Note that attempts have been made to do this by the use of dummy variables (variables that take on the value of zero for bombers and one for the fighters). In the linear regression model, this technique will work if the strata have either common slopes or intercepts. Unfortunately this will not be the case in general.

Independence. This assumption is also violated quite often. For instance, initial aircraft types and follow-on aircraft types are sometimes included in the same sample. Knowing the value of the initial aircraft type must change the distribution on the follow-on aircraft.

Random Sample. When building a CER we generally take all possible experience into the sample. Is this collection of data a sample or is it the population? If it is really a sample from a larger population, i.e., other sample points have not yet been built but theoretically exist, we still have the problem of whether or not the sample is random. A whole paper can probably be addressed to this question.

From the above, it becomes evident that the assumptions are generally violated in the cost analysis application. Of course in each application one strives to satisfy as many of the assumptions as possible. But, we often fail.

Unfortunately, the usual interpretation of many of the statistics and interval estimates that we wish to make use of depend heavily on the assumptions. Should we then throw the whole tool away? No. The statistics are still a valuable aid, and the extent to which they can be used and the interpretation that can be given to them will become clearer as we proceed.

SELECTION OF A CRITERION OF "BEST"

Before discussing particular statistics, a few remarks should be made about why the estimates of the parameters a, $b_1, \ldots, b_k$ are picked to minimize $\sum (Y_i - (a + b_1 X_{1i} + \ldots + b_k X_{ki}))^2$, that is, why least squares estimators are picked as best estimators.

First the method is intuitively appealing. One way to judge how well a particular CER will predict is to see how well it fits past observations. Therefore, picking the estimates of the parameters to minimize some additive form of the differences between the observed and what would be predicted by the CER is desirable. Of course the above expression is only one of a number of functional forms that will do this.

Secondly, the method is mathematically convenient. To understand how important this property is one need only look for examples of cases where non-linear regression functions are hypothesized. With the exception of the log-linear functional form or some other functional form that can be transformed into a linear model, one would have a great deal of difficulty finding such examples. Even in the log-linear example, the procedure usually consists of taking the logarithms, so that a linear form is obtained.[1] Hence, estimators of the parameters are picked to minimize $\sum (\ln Y_i - \ln (a X_{1i}^{b_1} X_{2i}^{b_2} \ldots X_{ki}^{b_k}))^2$. In effect best estimators

[1] A routine is available at RAND that uses an iterative technique to solve for the parameters without taking logarithms. See Boren, H. E. and Graver, C. A., Multivariate Logarithmic and Exponential Regression Models, RM-4879 PR, RAND Corporation, 1966.

are therefore defined differently (as can be seen by comparing the two summations being minimized), so that a solution for the estimators can be obtained. Note that the statistical assumptions are usually made in the linear form so the usual statistics are obtained for the ln Y_i and not Y_i itself. Hence, these statistics are not directly comparable to those obtained for Y_i in a linear model.

Thi. dly, the estimators are unbiased, that is $E\hat{a} = a$ and $E\hat{b}_p = b_p$, where E stands for expected value of. This is a nice property, but not a necessary one. In fact some biased estimators have less variance than their unbiased counterparts. However, it sounds un-American to have a biased estimator, so we had better stick with unbiased ones.

The estimators also have a minimum variance property. By the Gauss-Markoff theorem,[2] the least squares estimators have simultaneously smaller variances among the class of linear, unbiased estimators. This, of course, is highly desirable if you want linear, unbiased estimators.

Finally, the least squares estimators are the Maximum Likelihood Estimators. This means that the particular parameters have been picked that maximize the probability content of the sample among the class of regression functions (represented by different values of a, $b_1, \ldots, b_k$) being considered. This is portrayed graphically in Figure 2 for the simple linear regression case. The same sample is

[2] Lindgren, B. W., Statistical Theory, Macmillan, New York, 1962, page 387.

represented in both graphs. The line in the left graph, however, is a poor choice because most of the points lie in the tail of the density functions and hence the sample has a low probability content. The line in the right graph contains a higher probability content for the sample as the points generally lie under the bell of the density function. This line is closer to the Maximum Likelihood Estimate.

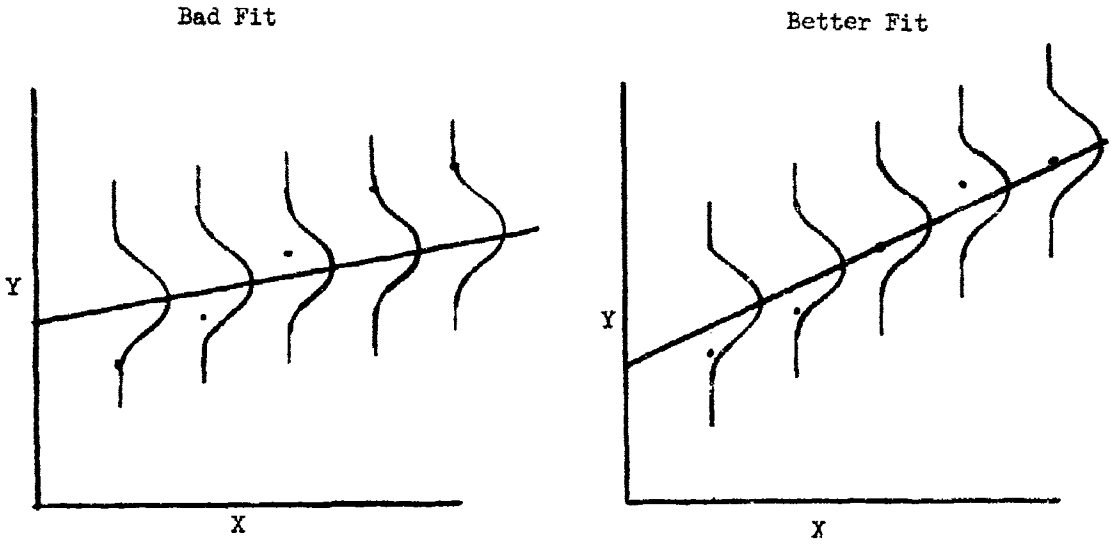


Figure 2

It is apparent from the above that the choice of least squares estimators as "best" has a lot going for it. Note that the first two properties listed have nothing to do with the statistical assumptions. Hence, even if the statistical assumptions are violated, the least squares technique can be used to obtain an estimating relationship. The process should probably be labeled curve fitting but whatever label is put on it, it still is of value as long as fitting the observed data is the basis for judging the goodness of the cost estimating relationships.

COMMONLY USED STATISTICS

Having decided to use least square estimators and after obtaining the estimates, the analyst is usually interested in answering questions like the following:

1. How well does the relationship fit the sample experience?

2. How does the particular form of the regression function compare with other pre-selected forms of the regression function?

Convenient tools for answering these questions are found in the following statistics:

1. Standard Error of the Estimate: (S_Y)

2. Correlation Coefficient: (r)

3. "F" Statistics

It turns out that all these statistics are related. Their relationship can be seen in a geometrical model.[1] Examining this model will hopefully add some insight into the behavior of the statistics as well as indicate how they may be validly used when the regression theory assumptions are not fulfilled.

As in any mathematical discussion, a number of definitions will have to be made for notational convenience. Three models will be looked at.

[1] This model can be found in most advanced statistics books that treat the topic of regression theory. In particular, it can be found in Lehmann, E., Testing Statistical Hypotheses, Wiley, New York, 1952.

In the one dimensional model it is assumed that the regression function has the form $Y=a$, a constant. That is each observation will be estimated by the same constant. The least squares estimate of the parameter is $\bar{Y}$, the arithmetic average of the observations in the sample.

The two dimensional model assumes a regression function of the form $Y = a + b_1 X_1$. The least squares estimates of a and b are denoted by $\hat{a}$ and $\hat{b}_1$, and the least squares estimate of Y is then denoted by $\hat{Y}$ and is equal to $\hat{a} + \hat{b}_1 X$.

Similarly, the three dimensional model has $Y = a + b_1 X_1 + b_2 X_2$ as the form of the regression function. $\tilde{Y}$ is the least squares estimate and is equal to $\tilde{a} + \tilde{b}_1 X_1 + \tilde{b}_2 X_2$ where $\tilde{a}, \tilde{b}_1, \tilde{b}_2$ are the least squares estimates of the three parameters.

The models are related in the sense that X_{1i} must represent the same characteristic or combination of characteristics in both the two and three dimensional models. Thus, for an airframe estimating problem, we might have the following:

Y_i = cost of the ith airplane

X_{1i} = weight times speed squared for the ith airplane

and X_{2i} = thrust of the ith airplane.

The models are summarized below:

Dimension	Regression Function	Least-Squares Estimate
One	$Y_i = a$	$a = \bar{Y}$
Two	$Y_i = a + b_1 X_{1i}$	$\hat{Y}_i = \hat{a} + \hat{b}_1 X_{1i}$
Three	$Y_i = a + b_1 X_{1i} + b_2 X_{2i}$	$\tilde{Y}_i = \tilde{a} + \tilde{b}_1 X_{1i} + \tilde{b}_2 X_{2i}$

There are various sums of squared deviations that are usually of interest. These are defined below:

Sums of Squared Deviations

Type	Definition	Notation
Total	$\sum (Y_i - \bar{Y})^2$	T^2
Explained by two dimensional model	$\sum (\hat{Y}_i - \bar{Y})^2$	E^2
Unexplained by two dimensional model	$\sum (Y_i - \hat{Y}_i)^2$	U^2
Explained by three dimensional model over the one dimensional model	$\sum (\tilde{Y}_i - \bar{Y})^2$	$\tilde{E}^2$
Unexplained by three dimensional model	$\sum (Y_i - \tilde{Y}_i)^2$	$\tilde{U}^2$
Additional explained by three dimensional model over two dimensional model	$\sum (\tilde{Y}_i - \hat{Y}_i)^2$	$\tilde{E}_A^2$

Notice that each of these sum of squares represents the square of the Euclidean distance between points in n dimensional space, the sample space. This fact is the basis of the geometrical model. The points I refer to are $(Y_1, Y_2, \ldots, Y_n)$, the observed sample; $(\bar{Y}, \bar{Y}, \ldots, \bar{Y})$, the "best" one dimensional model explanation of the observed sample; $(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n)$, the "best" two dimensional model explanation of the observed sample; and $(\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_n)$, the "best" three dimensional model explanation of the observed sample.

An initial view of the model is presented in Figure 3. Each point in the space represents one possible outcome of the sample of observations, $(Y_1, \ldots, Y_n)$. Two points have been identified, the origin and the sample we have observed.

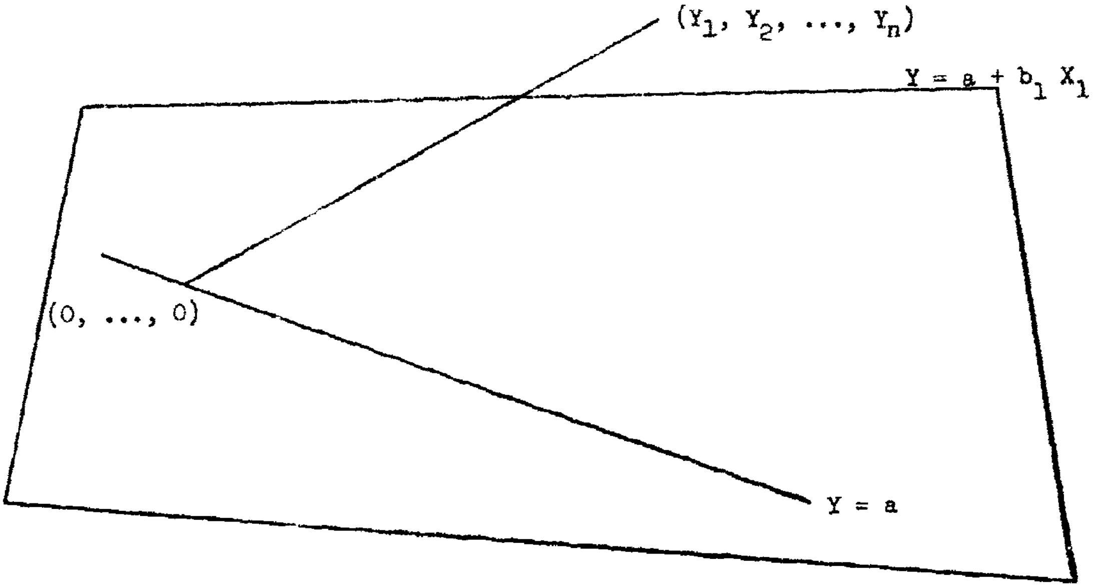
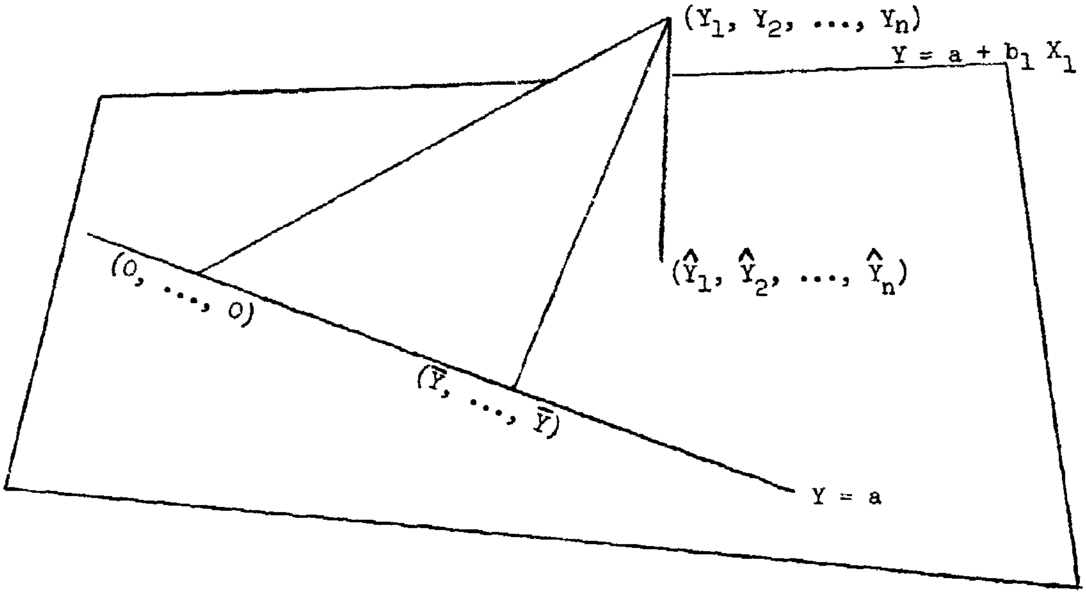


Figure 3

The line $Y = a$ is a representative of all the possible one dimensional models, while the plane $Y = a + b_1 X_1$ represents all possible two dimensional models. That is each point on the plane represents a particular value for "a" and "b_1". The line, $Y = a$, lies in the plane and represents those models for which $b_1 = 0$.

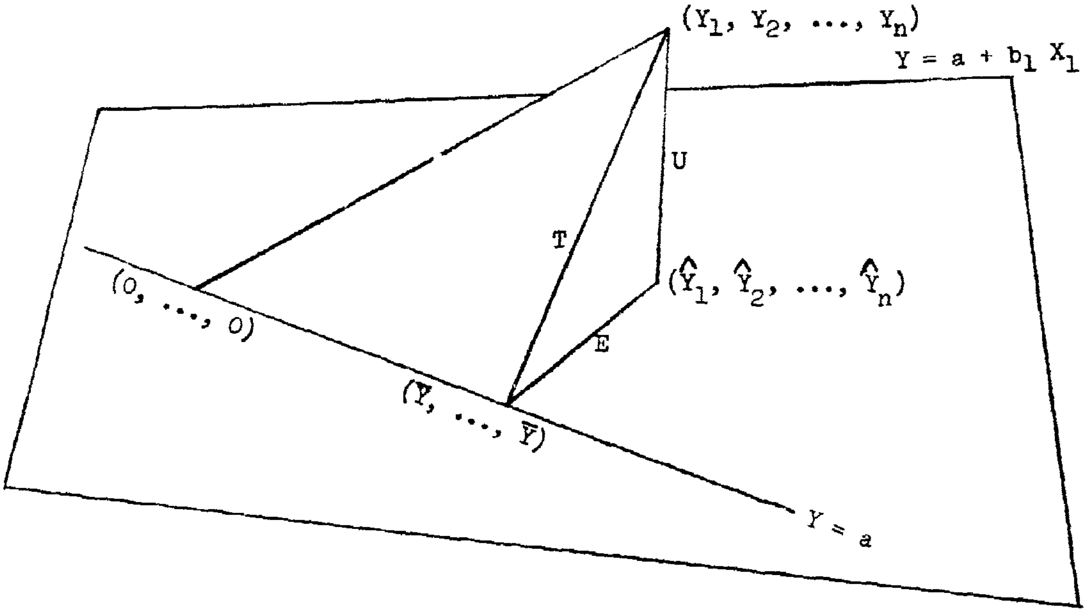
If we orthogonally project the sample point on to the line and the plane (Figure 4), the least squares estimate of the parameters are obtained, that is, $a = \bar{Y}$ for the projection onto the line, and $a = \hat{a}$, $b_1 = \hat{b}_1$ for the projection on to the plane. This is because a line connecting the orthogonal projection to the sample point is perpindicular to the plane (line) and hence is the shortest line between the point and the plane (line). But the square of the Euclidean distance of a line from the sample point to any point on the plane (line) is nothing more than the sum of squared deviations about the point on the plane. For example, suppose the point was given by $\dot{Y}_i = \dot{a} + \dot{b}_1 X_{1i}$. Then the square of the Euclidean distance is given by $\sum (Y_i - \dot{Y}_i)^2$. In the orthogonal projection, the shortest line is picked and hence this quantity is minimized. But that is equivalent to the least squares estimating procedure, i.e., we pick $\hat{a}$ and $\hat{b}_1$ so that the sum of the squared deviations is minimized.



$(Y_1, Y_2, \ldots, Y_n)$
$Y = a + b_1 X_1$
$(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n)$
$(0, \ldots, 0)$
$(\bar{Y}, \ldots, \bar{Y})$
$Y = a$

Figure 4

Now complete the triangle by drawing a line between $(\hat{Y}_1, \ldots, \hat{Y}_n)$ and $(\bar{Y}, \ldots, \bar{Y})$. This triangle (see Figure 5) is a right triangle (because of the orthogonal projection) and the length of the sides are given by T, E and U. The square of the distance of the sides are nothing more than the total sum of squares, explained sum of squares and unexplained sum of squares, defined earlier. Of course $T^2 = E^2 + U^2$, a consequence of the Pythagorean theorem.



$(Y_1, Y_2, \ldots, Y_n)$
$Y = a + b_1 X_1$
U
T
$(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n)$
E
$(0, \ldots, 0)$
$(\bar{Y}, \ldots, \bar{Y})$
$Y = a$

Figure 5

What about the statistics I referred to earlier? How do they fit into the model? To begin with the standard error of the estimate, denoted by S_Y, is equal to $U/\sqrt{n-2}$. The correlation coefficient, r, is equal to E/T. The F test for $b_1 = 0$ is given by $\frac{E^2}{U^2/(n-2)}$. The latter has the F distribution since, under the assumption that Y = a is the true model, T^2, E^2 and U^2 have chi-square distributions with n-1, 1 and n-2 degrees of freedom, respectively. Furthermore U^2 and E^2 are independent (expressed geometrically by the right angle between sides E and U).

So the common statistics that are used are no more than comparisons of distance (sometimes adjusted for degrees of freedom). Let's see how they work. When $(Y_1, \ldots, Y_n)$ is close to the plane, i.e., the sample observations fit the two dimensional model well, then U is small and E is almost as large as T. Hence, the standard error of the estimate is small, the correlation coefficient is close to 1 and F is large (significant).

Of course when the sample point is far away from the plane, i.e., it does not fit the two dimensional model (using X_1) so well, then U is large and E is small when compared with T so that S_Y is large, r is close to 0 and F is small (insignificant). Notice that the relationship between the point and the plane depends on X_1. The fact that this two dimensional model is insignificant does not rule out all two dimensional models. There may be a different definition of X_1, for example speed instead of weight times speed squared, that will describe a plane that lies closer to the observed sample.

It should be noted that the F statistic used in this model is the square of the usual "t" statistic that is used to test $b_1 = 0$. Either statistic can be used, but the F statistic is more general and can be used for comparing models of different dimension such as three dimensions with one dimension.

Now let's add another dimension to the model. To retain the correlation coefficient it will be necessary to start with $(\bar{Y}, \ldots, \bar{Y})$ as the origin. This is because we are limited to the number of dimensions that can be portrayed in a two dimensional picture. Shown in this model (Figure 6) is the same triangle that was examined earlier, but now the line represents what the plane represented in the previous picture, that is we ha e a two dimensional line.

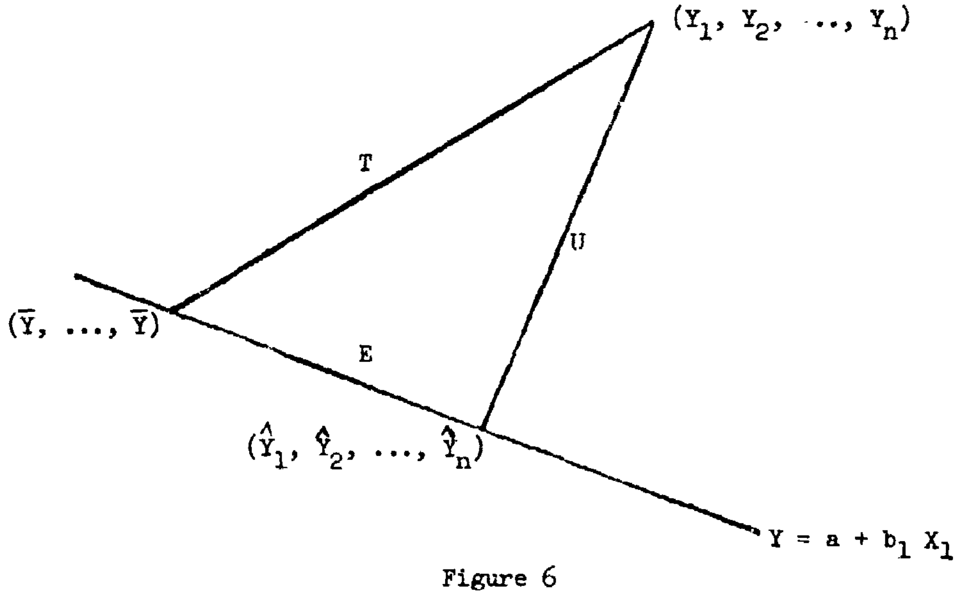

$(Y_1, Y_2, \ldots, Y_n)$
T
U
$(\bar{Y}, \ldots, \bar{Y})$
E
$(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n)$
$Y = a + b_1 X_1$

Figure 6

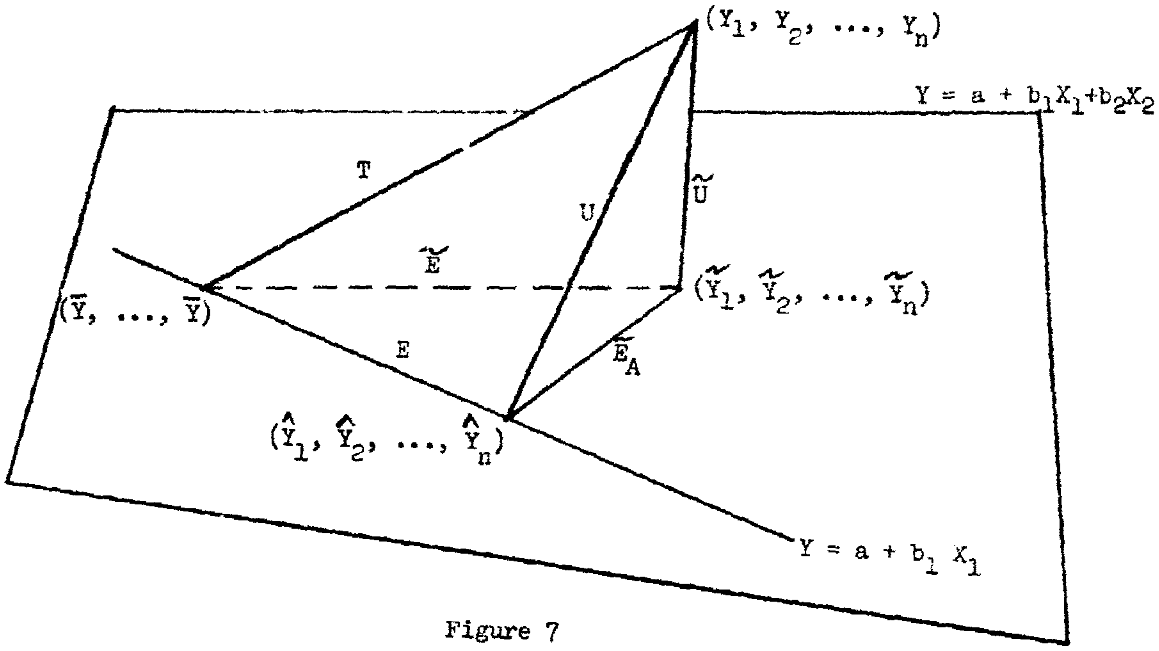

$(Y_1, Y_2, \ldots, Y_n)$
$Y = a + b_1X_1+b_2X_2$
T
U
$\tilde{U}$
$\tilde{E}$
$(\bar{Y}, \ldots, \bar{Y})$
$(\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_n)$
E
$\tilde{E}_A$
$(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n)$
$Y = a + b_1 X_1$

Figure 7

The three dimensional model (Figure 7) can now be added in the form of a (3 dimensional) plane. As before, the sample point can be orthogonally projected onto the plane to pick up the least squares estimates $(\tilde{Y}_1, \ldots, \tilde{Y}_n)$. And we can orthogonally project that point onto the 2 and 1 dimensional models, completing two new triangles. The squares of the new sides of the triangles obtained represent the following:

$\tilde{E}^2$ is the sum of squares explained over the one dimensional model

$\tilde{E}_A^2$ is the sum of squares explained over the two dimensional model

and $\tilde{U}^2$ is the still unexplained sum of squares.

As before we have the statistics for the two dimensional models:

$$S_Y = U/\sqrt{n-2}$$

$$r = E/T$$

$$F = \frac{E^2}{U^2/(n-2)} \quad \text{which tests } b_1 = 0$$

But now we also have the comparable statistics for the three dimensional model. These are given by:

$$S_Y = \tilde{U} / \sqrt{n-3}$$

$$r = \tilde{E}/T$$

$$\text{and} \quad F = \begin{cases} \dfrac{\tilde{E}^2/2}{\tilde{U}^2/(n-3)} & \text{Tests } b_1 = b_2 = 0 \\ \dfrac{\tilde{E}_A^2}{\tilde{U}^2/(n-3)} & \text{Tests } b_2 = 0 \end{cases}$$

Two interesting facts come out of this picture. First the triangle $\tilde{E}$, $\tilde{U}$, T is used to compare the three dimensional model to the one dimensional model while the triangle U, $\tilde{E}_A$, $\tilde{U}$ is used to compare the three dimensional model to the two dimensional model. In the former case we have have the F statistic that simultaneously tests $b_1 = b_2 = 0$. In the latter we test $b_2 = 0$ only. Notice that the F statistics are adjusted for degrees of freedom. There is a penalty that must be paid for going to a model of higher dimension, as the more dimensions you have the easier it is to explain the sample point by chance.

Secondly, the correlation coefficient, r, is larger in the three dimensional model than the two. This can be seen by the following:

$$r_{3\text{dim}} = \frac{\tilde{E}}{T} = \frac{\sqrt{E^2 + \tilde{E}_A^2}}{T} \geq \frac{E}{T} = r_{2\text{dim}}$$

Each new dimension that is added will explain more of T. Thus if r is being used as the criterion of "best." the higher dimensional model will always be preferred, even if the F test for $b_2 = 0$ is insignificant. This is why some people adjust r for degrees of freedom.

It should be pointed out at this time that we are not limited in this geometrical model to comparing 3 dimensional, 2 dimensional and 1 dimensional models. The plane could represent a k dimensional model while the line could represent a p dimensional model. All that is required is that $k > p$, and $X_1, X_2, \ldots, X_p$ be identical for the two models. Of course S_Y and the F tests would have to be adjusted for the correct degrees of freedom and the F test between k and p dimensions would test $b_{p+1} = b_{p+2} = \ldots = b_k = 0$ while the F test between one dimension and k dimensions would test all b's equal to zero. The degrees of freedom are easy to figure out. They are equal to the difference in the dimension of the models that the line is connecting. T^2 connects the sample point in n space with the point representing the best one dimensional model. Hence, T^2 has n-1 degrees of freedom. Similarly the following lines have the identified degrees of freedom.

U^2 has n-p
E^2 has p-1
$\tilde{U}^2$ has n-k
$\tilde{E}^2$ has k-1
and $\tilde{E}_A^2$ has k-p

An interesting question can be raised at this point. Why do we use F tests, rather than (adjusted) r to compare different models? One important factor is that the distribution is known. But even more important it has been shown that statistically they are the best tests available. This is a consequence of the Neyman-Pearson Lemma[2] which says the following:

> For any given Probability of rejecting the 2 dimensional model when it is true, the probability of rejecting the 3 dimensional model when it is true is minimized. Statisticians have labeled this property most powerful.

By using this model one can discover the reasons for the behavior of the statistics and guard oneself against certain pitfalls. An example of this is pictured in Figure 8. In this example if one just looked at the three dimensional model, he would be very pleased. S_Y (equal to $\tilde{U}/\sqrt{n-3}$) is small, r (equal to $\tilde{E}/T$) is close to one and F (equal to $\frac{\tilde{E}^2/2}{\tilde{U}^2/(n-3)}$) is significant. But looking at the two dimensional model S_Y (equal to $U/\sqrt{n-2}$) is small, r (equal to E/T) is close to one and F (equal to $\frac{E^2}{U^2/(n-2)}$) is significant also. The reason for this, of course, is that the three dimensional model is insignificant when compared to the two dimensional model, that is, F (equal to $\frac{\tilde{E}_A^2}{\tilde{U}^2/(n-3)}$) is small. Hence, from the statistical point of view, the two dimensional model is preferred.

[2] See Lindgren, B. W., Statistical Theory, Macmillan, New York, 1962, page 238.

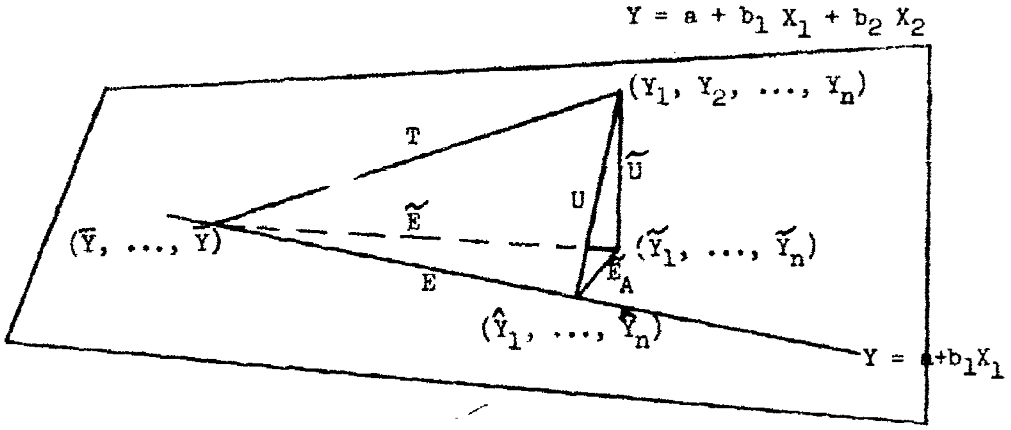

$Y = a + b_1 X_1 + b_2 X_2$
$(Y_1, Y_2, \ldots, Y_n)$
T
$\tilde{U}$
U
$\tilde{E}$
$(\bar{Y}, \ldots, \bar{Y})$
$(\tilde{Y}_1, \ldots, \tilde{Y}_n)$
E_A
E
$(\hat{Y}_1, \ldots, \hat{Y}_n)$
$Y = a+b_1X_1$

Figure 8

Notice it is from the statistical point of view that the two dimensional model would be preferred. If all of the assumptions mentioned earlier were satisfied this would be the case. But as was pointed out earlier, the assumptions are usually not satisfied. How do we make use of all these statistics then?

If it is agreed that it is desirable to fit the sample observations closely without increasing the number of independent variables substantially and if one keeps in mind that all of the statistics discussed are merely combinations of certain distances that describe how well the model fits the sample observations, then the statistics can be meaningfully used. As an example, suppose one has no reason for prefering one model over another. He might use an F test to decide between them. After all the test does provide a decision rule and the statistic is adjusted for degrees of freedom, thus penalizing the higher dimensional model for some of the "better fit" implied in the technique. In using this test, however, one should not make a level of significance statement (which is a statistical statement that depends on the assumptions), and one should display the value of the F statistic so that a user would be able to judge if the conclusion is consistent with his own model preferences.

Suppose now that one does have a strong preference for one of the two models (based perhaps on some physical relationship). Then he can adjust the decision point (the value that divides the acceptance region of the lower dimensional model from the rejection region) to a higher value if the lower dimensional model is preferred or to a lower value if the higher dimensional model is preferred. This would probably be done implicitly and as in the no preference case, the value of the F statistic should be portrayed so that the user will be able to apply his own model preferences. This use of the F statistic will put one in the position of rejecting "statistically" significant terms or accepting "statistically" insignificant terms. However, the point is that if the assumptions are not satisfied the decision points described in the tables for the F test are no more meaningful than a decision rule (such as value of the F statistic) supplied by the analyst.

It should be pointed out that care should be used in retaining a term that has a small value for the F statistic, for this is an indication that the sample contains little information concerning the value of the coefficient of this term. In this case one ought to consider somehow independently picking the value of this coefficient, i.e., not using the least squares estimate of the coefficient. The right triangle in the model will be lost in this case, but the advantages could outweigh this disadvantage.

It should also be pointed out that during the preceding discussion, the F statistic has played a dominate role. If the statistical assumptions were satisfied, then this domination is justified (see bottom of page 24). But in the cost analysis application, the favoritism of the F test cannot be championed as strongly. The correlation coefficient (preferably adjusted for degrees of freedom) can justifiably be used as a decision rule and the standard error of the estimate can provide valuable information concerning the amount of unexplained variance. Other measures, such as the coefficient of variation[3] which have not been discussed in this paper but are functionally related to those measures discussed, can also supply meaningful information. The choice of what measures to use is not nearly as important as keeping in mind what the measures mean and displaying the value of the statistics (distances) so that another analyst can decide what measure to use and what decision rule to follow.

[3] Fisher, G. H., Use of Statistical Regression Analysis in Deriving Estimating Relationships; Concepts and Procedures of Cost Analysis; RM 3589-PR, RAND Corporation, June 1963, P. V-17.

INTERVAL ESTIMATES

So far we have been concerned with comparing multiple linear models, and have discussed the types of statistics that will help us to choose between them. Another group of statistics, interval estimates, is used to make statements about the range of values that a variable of interest may take.

A main reason for discussing interval estimates is to point out the similarities and differences between them. Very often the types of interval estimates are confused. The intervals to be discussed are the confidence interval, prediction interval,[1] and the interval based on the standard error of the estimate which I will label the standard error interval for lack of a better name.

A few notational symbols will be helpful. In this section the following definitions of symbols will hold:

Y : the random variable of interest

Y_T : the true value of Y

$\hat{Y}$: the estimated value of Y, i.e.,

$$\hat{Y} = \hat{a} + \hat{b}_1 X_1$$

EY : the true expected value of Y, i.e.,

$$EY = a + b_1 X_1$$

The intervals have some similarities. All of them take the same form, e.g., $P(L_B < Y < L_u) = .95$. The form states that the probability

[1] Lindgren, B. W., Statistical Theory, Macmillan, New York, 1962, page 371.

of the random interval, described by random lower and upper bounds (L_B and L_u), covers the value of interest (Y) with probability .95 (or some other amount). Notice that I did not say that the probability of the value of interest lying in the interval is .95. The interval is random and if 100 such intervals were constructed from 100 independent samples, the statement says that we would expect to see 95 of them cover the value of interest.

Another similarity is the form of the bounds. They all take the form of $\hat{Y} \pm t \sqrt{\text{some variance measure}}$ where t is the value of the t statistic at the level desired.

But here the similarities end. The standard error interval is a statement about Y_T, and is only valid (from a prediction point of view) if $\hat{Y} = EY$. That is, we must have picked the right parameters, i.e., $\hat{a} = a$ and $\hat{b}_1 = b_1$ in the simple linear case. The variance that is used in the bounds is the square of the standard error of the estimate, denoted VAR Y_T, which is the estimate of σ^2, the variance of the error term in the original model. This interval estimate has bounds that are parallel to the estimated line of means, $\hat{Y}$. The interval can be used to describe the sample data but it is theoretically useless because it assumes $\hat{Y} = EY$.

The next interval to be examined is the confidence interval. It is not a statement about Y_T, but a statement about EY. Since this is usually not the prediction problem of interest, it also is not a very useful interval. The variance used in the bounds is an estimate of the variance inherent in selecting the estimated line of means $\hat{Y}$ and is denoted VAR $\hat{Y}$. It is equal to (in the simple linear case) VAR $\hat{a}$ + X^2 VAR $\hat{b}$. The bounds are no longer parallel but take on a parabolic shape. The bounds are closest together when $X = \bar{X}$ the arithmetic average of X's used in the sample.

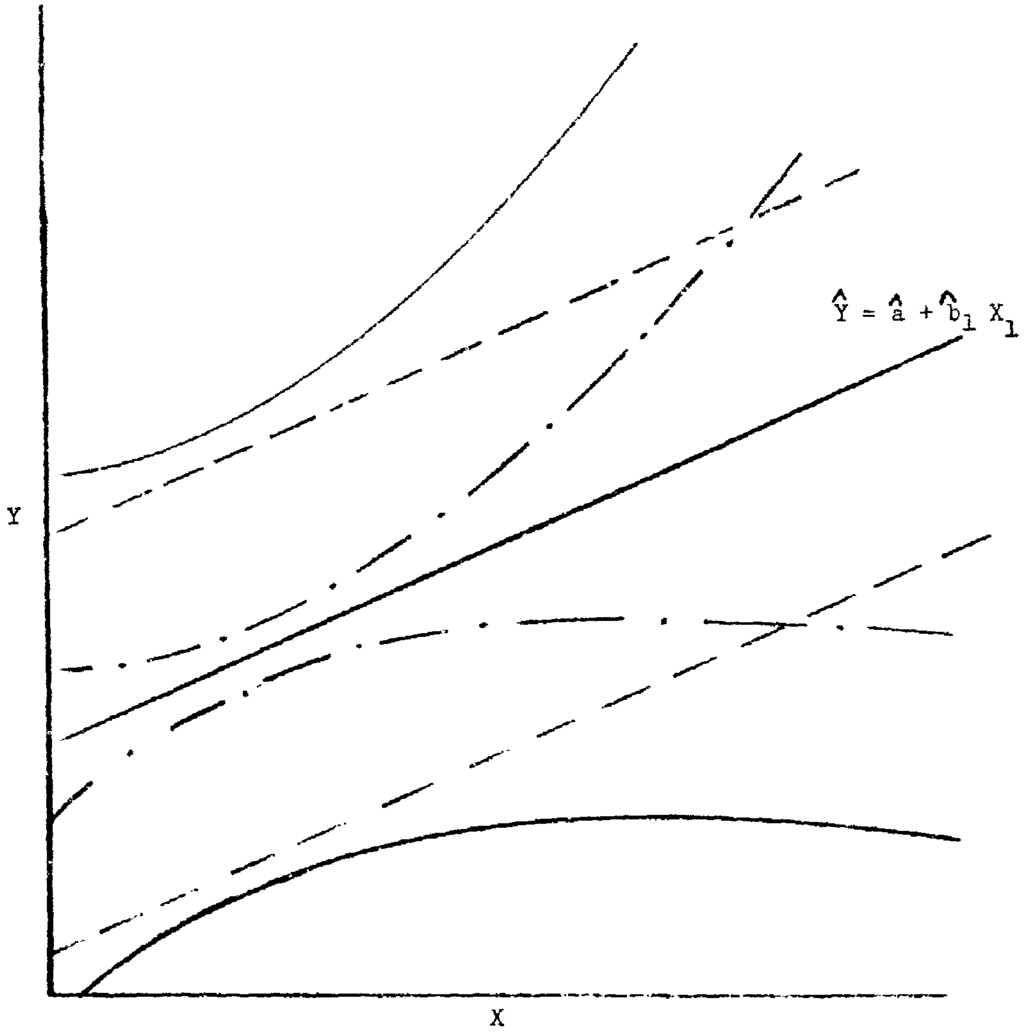
The last interval to be discussed is the prediction interval. It in a sense combines the Standard Error interval and the confidence interval. It is a statement about Y_T. The variance used in the bounds combines the variances previously discussed, hence the bounds pick up the estimate of the variances of e (the error term in the original model), $\hat{a}$ and $\hat{b}$. It is a statistically sound interval for prediction and hence is the most useful. The prediction interval is wider than either of the preceding intervals and like the confidence interval the bounds are parabolic in shape and closest together when $X = \bar{X}$.

A summary of the characteristics of the different intervals and a graphical representation of the usual relationship between the intervals are given below.

Characteristics

Interval	Interval On	Bounds
Standard Error	Y_T	$\hat{Y} \pm t \sqrt{\text{VAR } Y_T}$
Confidence	E_Y	$\hat{Y} \pm t \sqrt{\text{VAR } \hat{Y}}$
Prediction	Y_T	$\hat{Y} \pm t \sqrt{\text{VAR } Y_T + \text{VAR } \hat{Y}}$

USUAL RELATIONSHIP OF INTERVALS



Standard Error Interval - - -

Confidence Interval —— . ——

Prediction Interval ————

Figure 9

What good are these intervals in the Cost Analysis application? The prediction interval would be applicable provided all the assumptions that we discussed earlier were satisfied as it expresses precisely the statement we are interested in making in cost estimates. Namely with probability .95, the lower bound is less than Y_T is less than the upper bound. But we have already seen that these assumptions are often violated.

In the preceding section, we have seen how the F statistic could be used in comparing different linear models even though the assumptions were not satisfied. In a similar manner the prediction intervals might be useful in comparing models with different functional forms. In particular, they could be used in comparing linear regression functions with log-linear regression functions. The bounds can be compared (after the log-linear bounds have been exponentiated). The model with the narrowest bounds over the Y region of interest can be assumed to be the better model assuming that there is no reason to prefer one form of the regression function over the other. The intervals after all have taken into account the different variances that are working as well as the effect of the functional form of the regression on these variances.

But one must be careful in making such comparisons. As was pointed out earlier, the definition of best is different between the two models. What effect this has on the above comparison must be looked into. One might suggest that we use the same definition of best but this leads to some rather complicated mathematical problems and the solution for the estimators, if one exists, will probably have to be found with the aid of a computer.

In any case when the validity of the assumptions are suspect, one should never make the strong probability content statement that is implicit in the interval estimates. We can perhaps talk about the comparison of the .95 prediction intervals but the conclusion cannot be drawn that the interval covers Y_T with probability .95 or any other probability.

CONCLUSION

In conclusion, we have discussed the fact that in general the Cost Analysis application does not satisfy the assumptions of regression theory. Even so the cost analyst has a problem to solve. He must develop CERs and in so doing he must usually choose between different CER candidates. It has been shown that even if the assumptions are not satisfied, statistics such as F tests can be used to pick between various multiple linear regression models. Prediction intervals might be used to compare non linear regression functions with linear regression functions. The theoretical statistician might argue that we are fooling ourselves by using these techniques. But each of these techniques is based on determining how well the model fits the past data. Since a good fit of historical data is about all we have to go on in building our CERs, it would seem that the techniques discussed can be effectively used in a comparative fashion to provide a decision between models.